PLOS ONE

RESEARCH ARTICLE

A discussion on significance indices for contingency tables under small sample sizes

Natalia L. Oliveira[1☯], Carlos A. de B. Pereira[2☯], Marcio A. Diniz[3☯], Adriano Polpo[3☯]*

1 Department of Statistics and Data Science, Carnegie Mellon Univesity, Pittsburgh, United States of America, **2** Department of Statistics, University of Sao Paulo, Sao Paulo, Brazil, **3** Department of Statistics, Federal University of Sao Carlos, Sao Carlos, Brazil

☯ These authors contributed equally to this work.
* polpo@ufscar.br

Citation: Oliveira NL, Pereira CAdB, Diniz MA, Polpo A (2018) A discussion on significance indices for contingency tables under small sample sizes. PLoS ONE 13(8): e0199102. https://doi.org/10.1371/journal.pone.0199102

Editor: Mauro Gasparini, Politecnico di Torino, ITALY

Received: November 15, 2017

Accepted: May 31, 2018

Published: August 2, 2018

Data Availability Statement: The computer code that was used to generate all analysis in the paper is available from Github (https://github.com/adrianopolpo/contingencytables).

Funding: This work was partially supported by the Brazilian agencies FAPESP grant 2012/16669-4, and CNPq grants 302767/2017-7 and 308776/2014-3. The agencies had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Hypothesis testing in contingency tables is usually based on asymptotic results, thereby restricting its proper use to large samples. To study these tests in small samples, we consider the likelihood ratio test (LRT) and define an accurate index for the celebrated hypotheses of homogeneity, independence, and Hardy-Weinberg equilibrium. The aim is to understand the use of the asymptotic results of the frequentist Likelihood Ratio Test and the Bayesian FBST (Full Bayesian Significance Test) under small-sample scenarios. The proposed exact LRT p-value is used as a benchmark to understand the other indices. We perform analysis in different scenarios, considering different sample sizes and different table dimensions. The conditional Fisher's exact test for 2 × 2 tables and the Barnard's exact test are also discussed. The main message of this paper is that all indices have very similar behavior, except for Fisher and Barnard tests that has a discrete behavior. The most powerful test was the asymptotic p-value from the likelihood ratio test, suggesting that is a good alternative for small sample sizes.

Introduction

We discuss indices for homogeneity, independence, and Hardy-Weinberg equilibrium hypotheses [1, 2] in contingency tables. We propose an exact evaluation of the Likelihood Ratio Test (LRT) as a benchmark significance index. Based on the work of [3], its idea is to evaluate the probability distribution of all possible tables on the sample space under the null hypothesis. Once the distribution for sampling contingency tables under the hypothesis is known, we are able to compute the exact distribution of the Likelihood Ratio Test (LRT) statistics. The main difficulty for this procedure is that it is computationally time-consuming, being only feasible for small sample sizes and/or for tables of small dimension.

The exact LRT p-value presented as a way to do exact inference. The aim is to compare the behavior of the frequentist LRT asymptotic p-value [4], the exact LRT p-value, the Fisher's exact test p-value [5], the Chi-Square test asymptotic p-value [6, 7] and the Barnard's exact test p-value [8–11]. These frequentist indices are also compared to the e-value from the Full Bayesian Significance Test (FBST) [12, 13]. It was considered the asymptotic e-value and its

approximation (based on a Markov Chain Monte Carlo procedure) of the exact e-value. The choice of adding a Bayesian index to the comparison study originates from the known asymptotic relationship between the LRT and the FBST [14]. Moreover, the FBST and its e-value can be viewed as a Bayesian p-value counterpart, and therefore it is interesting to understand this Bayesian method when compared to frequentist methods. It is important to point out that we are mainly interested in the values of the indices, not in the acceptance or rejection of the hypothesis; that is, our focus is on the significance test, which consists of the evaluation of the p-(e-)values. In an applied setting, the researcher can, based on the indices, make his/her decision about his/her application. We are not interested in comparing the values of the indices with some fixed significance value (generally 5%) to decide the if the hypothesis should be accepted or rejected. With this goal in mind, all significance indices considered here are in agreement with the ASA's statement on significance indices [15].

From a historical perspective, hypothesis testing has been the most widely used statistical tool in many fields of science [16–18]. For categorical data, [19] discusses some exact procedures to perform inference and [20] presents methodological procedures for hypothesis testing for contingency tables. Tests for homogeneity hypothesis in contingency tables have been compared by [21], who compared the conditional and unconditional, and by [22], who compares, under an asymptotic perspective, two tests for equality of two proportions considering Goodman's Y^2 and χ^2 statistics. Regarding tests for the independence of two classifiers in contingency tables, [23] presents an algorithm for finding the exact permutation significance level for $r \times c$ contingency tables. [24], studies a simple way to compare two correlated proportions. More recently, [25] presents the exact likelihood ratio test for equality of two normal populations, and [26] discuss exact unconditional tests for homogeneity hypothesis in 2×2 tables.

One important aspect that differentiates the tests procedures is how each one deals with the elimination of the nuisance parameter. Basu [27] lists several methods but focuses on marginalization and conditioning. He defines *marginalization* as every procedure that replaces the observed sample x by the observed value of a suitable statistic $T(x) = t$. Therefore, instead of working with the original experiment $\mathcal{E}$ and data x, one should use the *marginal* experiment $\mathcal{E}_T$ and the recorded value $T(x)$ since the *marginal* statistical model would depend only on the parameter of interest. To justify these procedures, Basu adds that researchers usually recur to invariance or partial sufficiency arguments.

By *conditioning*, Basu defines methods of elimination that also consist of choosing a suitable statistic, but such that the conditional distribution of the observed sample, x, given the observed value of the statistic depends on the full parameter space only through the parameter of interest. Another commonly used approach that Basu describes is the one he calls *maximization*. In this case the nuisance parameter is eliminated from the risk function by some sort of maximization (or minimax) principle or directly from the likelihood, usually maximizing it with respect the nuisance parameters.

A final important strategy mentioned by Basu is the one he called *Bayesian solution*. In this case, one should derive the full posterior and integrate out the nuisance parameters, obtaining the posterior marginal distribution necessary to perform the required statistical inference. It is important to point out that the FBST does not follow this Bayesian strategy, since its evidence value is computed considering the full posterior. The proposed exact LRT p-value is based on the idea of integrating out the nuisance parameter, which is in some way related to Basu's *Bayesian solution* [26]. The methods for elimination of nuisance parameters, *maximization* and *Bayesian solution* can be considered as unconditional methods.

The Likelihood Ratio Test (LRT) asymptotic p-value [28], the Chi-Square test asymptotic p-value [29], Fisher's homogeneity exact test [29, 30], Barnard's exact test [8], and the Full Bayesian Significance Test (FBST) asymptotic and exact e-value [12, 13] are presented in detail

for the case of 2 × 2 contingency tables considering homogeneity hypothesis (Section 1.1). The theoretical results for homogeneity and independence hypotheses for tables of any dimension and Hardy-Weinberg equilibrium hypothesis are presented in sections 1.2, 1.3 and 1.4.

We study the relationship between indices in Section 2.1. [14] perform a similar study, however they consider continuous random variables using the e-value and the LRT p-value and show that these indices share an asymptotic relationship. In our case, the asymptotic LRT p-value, the exact LRT p-value and the Chi-Square p-value have similar behavior, including in small sample size scenarios. Both Fisher's exact test and Barnard's exact test have a discrete behavior for their p-values, being more clear for the Barnard's exact test p-value. All tests are unconditional tests, except for the Fisher one, that is a conditional test. It is important to draw attention to the fact that the present results are not based on a simulation study, we compute the indices for all possible tables in the sample space.

In addition to our focus on the study of significance indices, we also provide, for the frequentist indices, a study of the power functions to compare the tests considering the homogeneity hypothesis (2 × 2 tables) and Hardy-Weinberg equilibrium hypothesis (Section 2.2). The Fisher's exact test was the least powerful, followed by the Barnard's exact test, Chi-Square test, the exact LRT and the asymptotic LRT, the most powerful one. We did not evaluate the power function for the FBST; firstly, because it is not the aim of the Bayesian paradigm, and secondly, to do so, it would be necessary to define a decision rule for the FBST, which is not in the scope of this paper. We also note that, under the hull hypothesis, considering the significance level 5%, all frequentist indices achieved 5% rejection as expected.

1 Methods

1.1 Homogeneity test for 2 × 2 contingency tables

Let X_1 and X_2 be two random variables, representing the rows (1 and 2) of Table 1, x_{11} and x_{21} being their observed values, and $n_{1\cdot}$ and $n_{2\cdot}$ fixed sample sizes. Consider the distributions of X_1 as Binomial($n_{1\cdot}$, θ_{11}) and X_2 a Binomial($n_{2\cdot}$, θ_{21}) for describing the chances of a subject belong to category (column) C_1 in two distinct populations. Both populations are partitioned into two categories (columns) C_1 and C_2 and the objective is to test homogeneity among the two unknown population frequencies, $\boldsymbol{H}$: $\theta_{11} = \theta_{21} = \theta$. This hypothesis is geometrically represented by a diagonal line of the unit square.

The likelihood function is specified by

$$L(\theta_{11}, \theta_{21} \mid x_{11}, x_{21}, n_{1\cdot}, n_{2\cdot}) = \frac{n_{1\cdot}! n_{2\cdot}!}{x_{11}! x_{21}! x_{12}! x_{22}!} \theta_{11}^{x_{11}} \theta_{21}^{x_{21}} (1-\theta_{11})^{x_{12}} (1-\theta_{21})^{x_{22}}, \quad (1)$$

where $0 \leq \theta_{i1} \leq 1$, $i = 1, 2$. Under $\boldsymbol{H}$, the likelihood function simplifies to

$$L(\theta \mid x_{11}, x_{21}, n_{1\cdot}, n_{2\cdot}, \boldsymbol{H}) = \frac{n_{1\cdot}! n_{2\cdot}!}{x_{11}! x_{21}! x_{12}! x_{22}!} \theta^{x_{11}+x_{21}} (1-\theta)^{x_{12}+x_{22}}, 0 \leq \theta \leq 1, \quad (2)$$

Table 1. Contingency table 2 × 2.

row\column	1	2	total
1	x_{11}	x_{12}	$n_{1\cdot}$
2	x_{21}	x_{22}	$n_{2\cdot}$

$n_{i\cdot} = x_{i1} + x_{i2}$, $i = 1, 2$.

https://doi.org/10.1371/journal.pone.0199102.t001

and the LRT test statistics is:

$$\begin{aligned}\lambda(x_{11}, x_{21}) &= \frac{\sup_{\theta\in\Theta_{\boldsymbol{H}}} L(\theta_{11}, \theta_{21}|x_{11}, x_{21}, n_{1\cdot}, n_{2\cdot})}{\sup_{\theta\in\Theta} L(\theta_{11}, \theta_{21}|x_{11}, x_{21}, n_{1\cdot}, n_{2\cdot})} \\ &= \frac{\left(\frac{x_{11}+x_{21}}{n_{1\cdot}+n_{2\cdot}}\right)^{x_{11}+x_{21}}\left(\frac{x_{12}+x_{22}}{n_{1\cdot}+n_{2\cdot}}\right)^{x_{12}+x_{22}}}{\left(\frac{x_{11}}{n_{1\cdot}}\right)^{x_{11}}\left(\frac{x_{12}}{n_{1\cdot}}\right)^{x_{12}}\left(\frac{x_{21}}{n_{2\cdot}}\right)^{x_{21}}\left(\frac{x_{22}}{n_{2\cdot}}\right)^{x_{22}}},\end{aligned} \tag{3}$$

in which $\Theta_{\boldsymbol{H}}$ is the parametric set defined by the hypothesis.

• **Exact LRT p-value:**

To define this p-value, we use the predictive distributions of X_1 and X_2 before any data were observed. The proposed p-value is an alternative way to calculate an exact p-value for the LRT. The goal is to find a distribution for the contingency table under $\boldsymbol{H}$ that is not a function on θ. We consider θ a nuisance parameter in the likelihood function in (2) and integrate it over θ in order to eliminate it, as suggested by [27]. The idea is to incorporate the concept of the *Bayesian solution* nuisance parameter elimination approach but in a frequentist setting, which means using the likelihood function instead of a posterior distribution. That is,

$$\begin{aligned}h(x_{11}, x_{21}) &= \int_0^1 L(\theta|x_{11}, x_{21}, n_{1\cdot}, n_{2\cdot}, \boldsymbol{H})\mathrm{d}\theta \\ &= \frac{n_{1\cdot}!n_{2\cdot}!}{x_{11}!x_{21}!x_{12}!x_{22}!}\int_0^1 \theta^{x_{11}+x_{21}}(1-\theta)^{x_{12}+x_{22}}\mathrm{d}\theta \\ &= \binom{n_{1\cdot}}{x_{11}}\binom{n_{2\cdot}}{x_{21}}\frac{(x_{11}+x_{21})!(x_{12}+x_{22})!}{(n_{1\cdot}+n_{2\cdot}+1)!} \\ &= \frac{\binom{n_{1\cdot}}{x_{11}}\binom{n_{2\cdot}}{x_{21}}}{\binom{n_{1\cdot}+n_{2\cdot}}{x_{11}+x_{21}}}\frac{1}{(n_{1\cdot}+n_{2\cdot}+1)}.\end{aligned} \tag{4}$$

To obtain the probability function $\Pr(X_1 = x_{11}, X_2 = x_{21} \mid \boldsymbol{H})$, one needs to find a normalization constant.

$$\Pr(X_1 = x_{11}, X_2 = x_{21} \mid \boldsymbol{H}) = \frac{h(x_{11}, x_{21})}{\sum_{i=0}^{n_{1\cdot}}\sum_{j=0}^{n_{2\cdot}} h(i,j)}. \tag{5}$$

Note that to calculate (5), we evaluate $h(\cdot, \cdot)$ for all possible tables. In the case of a homogeneity hypothesis for 2×2 contingency tables, $\sum_{i=0}^{n_{1\cdot}}\sum_{j=0}^{n_{2\cdot}} h(i,j) = 1$. We present the table's probability in terms of this sum to obtain a general formula for all hypotheses and table dimensions considered here, since in other scenarios this quantity does not sum up to 1 (for example, the sum of h for all possible 2×2 tables considering independence hypothesis with $n = 2$ is 2304). The exact p-value calculation follows directly from the test statistic distribution:

$$\begin{aligned}\text{p-value} &= \Pr(\lambda(X_1, X_2) \leq \lambda(x_{11}, x_{21}) \mid \boldsymbol{H}) \\ &= \sum_{(i,j)\in\boldsymbol{R}} \Pr(X_1 = i, X_2 = j \mid \boldsymbol{H}),\end{aligned} \tag{6}$$

in which $\boldsymbol{R}$ is the set of all pairs (i, j) such that $\lambda(i, j) \leq \lambda(x_{11}, x_{21})$, and $\lambda(x_{11}, x_{21})$ is the observed test statistic, as in (3).

• **Barnard's Exact Test:**

Consider that $n_{1\cdot}$ and $n_{2\cdot}$ are fixed in Table 1. The random variables X_1 and X_2 are independent Binomial distribution with parameters θ_{11} and θ_{21}. The probability of a sample $\{x_{11}, x_{21}\}$ be drawn is

$$\Pr(X_1 = x_{11}, X_2 = x_{21}) = \frac{n_{1\cdot}!}{x_{11}!x_{12}!}\theta_{11}^{x_{11}}(1-\theta_{11})^{x_{12}}\frac{n_{2\cdot}!}{x_{21}!x_{22}!}\theta_{21}^{x_{21}}(1-\theta_{21})^{x_{22}}, \tag{7}$$

and, under hypothesis $\boldsymbol{H}$,

$$\Pr(X_1 = x_{11}, X_2 = x_{21} \mid \boldsymbol{H}) = \frac{n_{1\cdot}!n_{2\cdot}!}{x_{11}!x_{12}!x_{21}!x_{22}!}\theta^{x_{11}+x_{12}}(1-\theta)^{x_{12}+x_{22}}. \tag{8}$$

We define the critical region as $\boldsymbol{R} = \{\lambda(X_1, X_2) \leq \lambda(x_{11}, x_{21})\}$, then the Barnard's exact p-value is obtained by

$$\text{p-value} = \max_{0\leq\theta\leq 1}\sum_{\boldsymbol{R}}\frac{n_{1\cdot}!n_{2\cdot}!}{x_{11}!x_{12}!x_{21}!x_{22}!}\theta^{x_{11}+x_{12}}(1-\theta)^{x_{12}+x_{22}}. \tag{9}$$

That is, the Barnard's exact test consider the p-values for all possible points of the parameter space under $\boldsymbol{H}$, and takes the maximum p-value. In this test, the chosen approach for nuisance parameter elimination among the ones presented by Basu is *maximization*.

• **Full Bayesian Significance Test:**

The Bayesian approach considered is based on the FBST (Full Bayesian Significance Test) [12, 13].

Definition 1 *Let* $\pi(\theta \mid \boldsymbol{x})$ *be the posterior density function of* θ *given the observed sample and* $T(\boldsymbol{x}) = \{\theta \in \Theta : \pi(\theta \mid \boldsymbol{x}) \geq \sup_{\theta\in\Theta_{\boldsymbol{H}}}\pi(\theta \mid \boldsymbol{x})\}$. *The supporting evidence measure for the hypothesis* $\theta \in \Theta_{\boldsymbol{H}}$ is defined as $Ev(\Theta_{\boldsymbol{H}}, \boldsymbol{x}) = 1 - \Pr(\theta \in T(\boldsymbol{x}) \mid \boldsymbol{x})$.

Consider that, *a priori*, θ_{11} and θ_{21} are independent and both follow a Uniform(0, 1) distribution. The choice of uniforms priors is to avoid a subjective prior to have a fair comparison with frequentist indices. Recall that X_1 and X_2 given θ_{11} and θ_{21} are Binomial distributed. Hence, the posterior distributions for θ_{11} and θ_{21} are independent Beta(x_{11} + 1, $n_{1\cdot} - x_{11}$ + 1) and Beta(x_{21} + 1, $n_{2\cdot} - x_{21}$ + 1). Under the hypothesis $\boldsymbol{H}$, the posterior distribution is

$$\pi(\theta \mid x_{11}, x_{21}, n_{1\cdot}, n_{2\cdot}, \boldsymbol{H}) = \frac{\theta^{x_{11}+x_{21}}(1-\theta)^{x_{12}+x_{22}}}{\mathcal{B}(x_{11}+1, x_{12}+1)\mathcal{B}(x_{21}+1, x_{22}+1)} \tag{10}$$

and by maximizing it in θ we obtain $\sup_{\theta\in(0,1)}\pi(\theta \mid x_{11}, x_{21}, n_{1\cdot}, n_{2\cdot}, \boldsymbol{H})$, where $\mathcal{B}(\cdot,\cdot)$ is the

Beta function. Since x_{11}, x_{21}, $n_{1\cdot}$ and $n_{2\cdot}$ are integers,

$$\pi(\theta \mid x_{11}, x_{21}, n_{1\cdot}, n_{2\cdot}, \boldsymbol{H}) = \binom{n_{1\cdot}}{x_{11}}\binom{n_{2\cdot}}{x_{21}}(n_{1\cdot}+1)(n_{2\cdot}+1)\theta^{x_{11}+x_{21}}(1-\theta)^{x_{12}+x_{22}}, \tag{11}$$

$$\sup_{\theta\in(0,1)} \pi(\theta \mid x_{11}, x_{21}, n_{1\cdot}, n_{2\cdot}, \boldsymbol{H}) = \frac{(n_{1\cdot}+1)!(n_{2\cdot}+1)!}{x_{11}!x_{21}!x_{12}!x_{22}!}\left(\frac{x_{11}+x_{21}}{n_{1\cdot}+n_{2\cdot}}\right)^{x_{11}+x_{21}}\left(\frac{x_{12}+x_{22}}{n_{1\cdot}+n_{2\cdot}}\right)^{x_{12}+x_{22}}, \tag{12}$$

the hypothesis' tangent set, T, is

$$T(x_{11}, x_{21}, n_{1\cdot}, n_{2\cdot}) = \left\{(\theta_{11}, \theta_{21}) \in (0,1)\times(0,1) : \pi(\theta_{11}, \theta_{21} \mid x_{11}, x_{21}, n_{1\cdot}, n_{2\cdot}) \geq \sup_{\theta\in(0,1)} \pi(\theta \mid x_{11}, x_{21}, n_{1\cdot}, n_{2\cdot}, \boldsymbol{H})\right\}, \tag{13}$$

and

$$\text{e-value} = 1 - \Pr[\theta \in T(x_{11}, x_{21}, n_{1\cdot}, n_{2\cdot})]. \tag{14}$$

To calculate the approximate e-value, we use the following algorithm:

1. A random sample of size k is generated from posterior distribution of θ_{11}, θ_{21}, obtaining $\{\theta_{x_{11}1}, \theta_{x_{21}1}\}, \ldots, \{\theta_{x_{11}k}, \theta_{x_{21}k}\}$.
2. The e-value is calculated by

$$1 - \frac{1}{k}\sum_{i=1}^{k} I\left(\pi(\theta_{x_{11}i}, \theta_{x_{21}i} \mid x_{11}, x_{21}, n_{1\cdot}, n_{2\cdot}) \geq \sup_{\theta\in(0,1)} \pi(\theta \mid x_{11}, x_{21}, n_{1\cdot}, n_{2\cdot})\right),$$

in which $I(A)$ is the indicator function of set A.

• Other indices:

For the LRT, the statistic $-2\ln[\lambda(X_1, X_2)]$ has asymptotically a chi-square distribution with 1 degree of freedom, which is $dim(\Theta) - dim(\Theta_{\boldsymbol{H}})$ [28]. The FBST uses the same statistic, however its asymptotic distribution is a chi-square with 2 degrees of freedom [13], which is $dim(\Theta)$. For the chi-square test and the Fisher's exact test for homogeneity see [29].

1.2 Homogeneity hypothesis for $\ell \times c$ contingency tables

Let X_i, $i = 1, \ldots, \ell$, be random variables that are represented by the rows of Table 2 and $n_{1\cdot}$, $n_{2\cdot}, \ldots, n_{\ell\cdot}$ are known constants.

Assuming that X_i, $i = 1, \ldots, \ell$, follows a Multinomial($n_{i\cdot}$, $\theta_{i1}, \ldots, \theta_{ic}$) distribution, we are interested in testing if their distributions are homogeneous with respect to categories

Table 2. Contingency table $\ell \times c$.

row\column	1	2	⋯	c	total
1	x_{11}	x_{12}		x_{1c}	$n_{1\cdot}$
2	x_{21}	x_{22}		x_{2c}	$n_{2\cdot}$
⋮			⋱	⋮	⋮
ℓ	$x_{\ell 1}$	$x_{\ell 2}$	⋯	$x_{\ell c}$	$n_{\ell\cdot}$
total	$n_{\cdot 1}$	$n_{\cdot 2}$	⋯	$n_{\cdot c}$	$n_{\cdot\cdot}$

$n_{i\cdot} = \sum_{j=1}^{c} x_{ij}$, $i = 1, \ldots, \ell$;

$n_{\cdot j} = \sum_{i=1}^{\ell} x_{ij}$, $j = 1, \ldots, c$; and

$n_{\cdot\cdot} = \sum_{i=1}^{\ell} \sum_{j=1}^{c} x_{ij}$.

https://doi.org/10.1371/journal.pone.0199102.t002

(columns) C_j, $j = 1, \ldots, c$. That is,

$$\boldsymbol{H}: \begin{cases} \theta_1 = \theta_{11} = \theta_{21} = \cdots = \theta_{\ell 1}, \\ \theta_2 = \theta_{12} = \theta_{22} = \cdots = \theta_{\ell 2}, \\ \quad \vdots \\ \theta_{c-1} = \theta_{1(c-1)} = \theta_{2(c-1)} = \cdots = \theta_{\ell(c-1)}, \end{cases}$$

in which $\theta_c = 1 - \sum_{k=1}^{c-1} \theta_k$, $0 \leq \theta_k \leq 1, \forall k = 1, \ldots, c$.

Let $\boldsymbol{x}$ be all observed values presented in Table 2 and $\boldsymbol{\theta}$ all the parameters. The likelihood function is

$$L(\boldsymbol{\theta} \mid \boldsymbol{x}) = \left[\prod_{i=1}^{\ell} n_{i\cdot}! \Big/ \prod_{i=1}^{\ell} \prod_{j=1}^{c} x_{ij}! \right] \prod_{i=1}^{\ell} \prod_{j=1}^{c} \theta_{ij}^{x_{ij}}, \tag{15}$$

and under the hypothesis $\boldsymbol{H}$,

$$L(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{H}) = \left[\prod_{i=1}^{\ell} n_{i\cdot}! \Big/ \prod_{i=1}^{\ell} \prod_{j=1}^{c} x_{ij}! \right] \prod_{j=1}^{c} \theta_j^{n_{\cdot j}}. \tag{16}$$

The LRT λ statistic is

$$\lambda(\boldsymbol{x}) = \prod_{j=1}^{c} \left(\frac{n_{\cdot j}}{n_{\cdot\cdot}} \right)^{n_{\cdot j}} \Big/ \prod_{i=1}^{\ell} \prod_{j=1}^{c} \left(\frac{x_{ij}}{n_{i\cdot}} \right)^{x_{ij}}. \tag{17}$$

• Exact LRT p-value:

To obtain the exact LRT p-value, we need the function $h(\boldsymbol{x})$. In this scenario,

$$h(\boldsymbol{x}) = \prod_{i=1}^{\ell} n_{i\cdot}! \prod_{j=1}^{c} n_{\cdot j}! \Big/ \left[\left(\prod_{i=1}^{\ell} \prod_{j=1}^{c} x_{ij}! \right) (n_{\cdot\cdot} + c - 1)! \right], \tag{18}$$

and the p-value's calculation follows as in Subsection 1.1.

• **FBST:**

Assuming a Dirichlet(1, 1, . . ., 1) prior for $\{\theta_{i1}, \ldots, \theta_{ic}\}$, and since $\boldsymbol{X_i}$ follows a Multinomial $(n_i, \theta_{i1}, \ldots, \theta_{ic})$ distribution, then the posterior distribution is a Dirichlet$(x_{i1} + 1, x_{i2} + 1, \ldots, x_{ic} + 1)$, $i = 1, \ldots, \ell$.

In this setting,

$$\sup_{\boldsymbol{\theta}\in\Theta_{\boldsymbol{H}}} \pi(\boldsymbol{\theta} \mid \boldsymbol{x}) = \prod_{i=1}^{\ell}(n_{i\cdot} + c - 1)! \prod_{j=1}^{c}\left(\frac{n_{\cdot j}}{n_{\cdot\cdot}}\right)^{n_{\cdot j}} \Big/ \prod_{i=1}^{\ell}\prod_{j=1}^{c} x_{ij}!, \tag{19}$$

and we can obtain the e-value from Definition 1.

• **Other indices:**

Both asymptotic LRT p-value and asymptotic e-value are calculated as $\Pr[-2\ln(\lambda(\boldsymbol{X})) \leq -2\ln(\lambda(\boldsymbol{x}))]$, but while the LRT considers that this statistic follows a X^2 distribution with $(\ell - 1)(c - 1)$ degrees of freedom, the FBST considers that it follows a X^2 distribution with $\ell(c - 1)$ degrees of freedom. The Chi-Square homogeneity test is also obtained.

1.3 Independence hypothesis for $\ell \times c$ contingency tables

Consider that θ_{ij} is the probability of observing a sample in the cell at row i and column j, $\theta_{i\cdot}$ is the probability of observing a sample in row i, $\theta_{\cdot j}$ is the probability of observing a sample in column j, $0 \leq \theta_{ij} \leq 1$, $0 \leq \theta_{i\cdot} \leq 1$, $0 \leq \theta_{\cdot j} \leq 1$, $i = 1, \ldots, \ell$, $j = 1, \ldots, c$, $\sum_{i=1}^{\ell}\sum_{j=1}^{c}\theta_{ij} = 1$, $\sum_{i=1}^{\ell}\theta_{i\cdot} = 1$, and $\sum_{j=1}^{c}\theta_{\cdot j} = 1$.

For the independence hypothesis, our interest is to test $\boldsymbol{H}$: $\theta_{ij} = \theta_{i\cdot} \times \theta_{\cdot j}$, $\forall i, j$. For the case of 2×2 table, the independence hypothesis is geometrically represented as Fig 1.

Considering that $n_{\cdot\cdot}$ is known, we assume that the outcomes of Table 2 follow a Multinomial$(n_{\cdot\cdot}, \boldsymbol{\theta})$ distribution, $\boldsymbol{\theta} = \{\theta_{11}, \ldots, \theta_{1(c-1)}, \ldots, \theta_{\ell 1}, \ldots, \theta_{\ell(c-1)}\}$, and $\theta_{ic} = 1 - \sum_{j=1}^{c-1}\theta_{ij}$, $i = 1, \ldots, \ell$. The likelihood function is

$$L(\boldsymbol{\theta} \mid \boldsymbol{x}) = \left[n_{\cdot\cdot}! \Big/ \prod_{i=1}^{\ell}\prod_{j=1}^{c} x_{ij}!\right] \prod_{i=1}^{\ell}\prod_{j=1}^{c}\theta_{ij}^{x_{ij}}. \tag{20}$$

The likelihood function under $\boldsymbol{H}$ is

$$L(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{H}) = \left[n_{\cdot\cdot}! \Big/ \prod_{i=1}^{\ell}\prod_{j=1}^{c} x_{ij}!\right] \prod_{i=1}^{\ell}\theta_{i\cdot}^{n_{i\cdot}} \prod_{j=1}^{c}\theta_{\cdot j}^{n_{\cdot j}}, \tag{21}$$

and the LRT λ statistic is

$$\lambda(\boldsymbol{x}) = \prod_{i=1}^{\ell}\left(\frac{n_{i\cdot}}{n_{\cdot\cdot}}\right)^{n_{i\cdot}} \prod_{j=1}^{c}\left(\frac{n_{\cdot j}}{n_{\cdot\cdot}}\right)^{n_{\cdot j}} \Big/ \prod_{i=1}^{\ell}\prod_{j=1}^{c}\left(\frac{x_{ij}}{n_{\cdot\cdot}}\right)^{x_{ij}}. \tag{22}$$

• **Exact LRT p-value:**

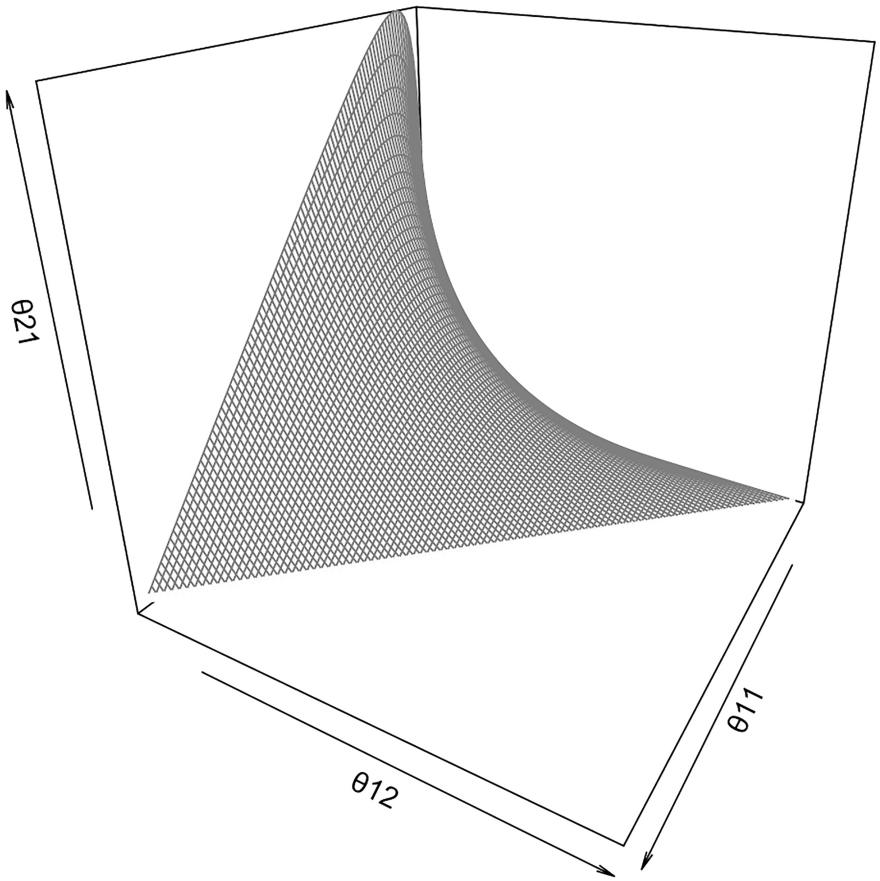


Fig 1. Geometric representation of the independence hypothesis (gray surface) for 2 × 2 tables. The parametric space is the three-dimensional simplex (regular tetrahedron).

https://doi.org/10.1371/journal.pone.0199102.g001

As shown in Subsection 1.1, this p-value is obtained the same way but with a different $h(\boldsymbol{x})$. In this case,

$$h(\boldsymbol{x}) = n_{..}!(n_{..}+\ell)!(n_{..}+c)!\Big/\left[\prod_{i=1}^{\ell}\prod_{j=1}^{c}x_{ij}!\prod_{i=1}^{\ell}n_{i\cdot}!\prod_{j=1}^{c}n_{\cdot j}!\right]. \tag{23}$$

• **FBST:**

Assuming a Dirichlet(1, . . ., 1) as prior distribution for $\boldsymbol{\theta}$ and that the outcomes of Table 2 follow a Multinomial(n, θ_{11}, . . ., θ_{1c}, . . ., $\theta_{\ell 1}$, . . ., $\theta_{\ell c}$) distribution, then the posterior distribution is a Dirichlet(x_{11} + 1, . . ., x_{1c} + 1, . . ., $x_{\ell 1}$ + 1, . . ., $x_{\ell c1}$ + 1). The e-value is obtained from Definition 1 and

$$\sup_{\boldsymbol{\theta}\in\Theta_H}\pi(\boldsymbol{\theta}\mid\boldsymbol{x}) = (n+\ell c-1)!\prod_{i=1}^{\ell}\left(\frac{n_{i\cdot}}{n}\right)^{n_{i\cdot}}\prod_{j=1}^{c}\left(\frac{n_{\cdot j}}{n}\right)^{n_{\cdot j}}\Big/\left[\prod_{i=1}^{\ell}\prod_{j=1}^{c}x_{ij}!\right]. \tag{24}$$

• **Other indices:**

We obtained the asymptotic LRT p-value and e-value, considering that $-2ln(\lambda(\boldsymbol{X}))$ follows a X^2 distribution with $(\ell - 1)(c - 1)$ and $(\ell c - 1)$ degrees of freedom. We also obtained the p-value for the Chi-Square independence test.

1.4 Hardy-Weinberg equilibrium

An individual's genotype is formed by a combination of alleles. If there are two possible alleles for one characteristic (say *A* and *a*), the possible genotypes are *AA*, *Aa* or *aa*. Considering a few premises true [31], the principle says that the allele probability in a population does not change from generation to generation. It is a fundamental principle for the Mendelian mating allelic model. If the probabilities of alleles are θ and $1 - \theta$, the expected genotype probabilities are $(\theta^2, 2\theta(1 - \theta), (1 - \theta)^2)$ $0 \leq \theta \leq 1$.

Considering the Hardy-Weinberg equilibrium, the aim is to verify if a population follows these genotypes proportions. Therefore, the equilibrium hypothesis is

$$\boldsymbol{H}: \begin{cases} \theta_1 &= \theta^2, \\ \theta_2 &= 2\theta(1-\theta), \\ \theta_3 &= (1-\theta)^2, \end{cases}$$

in which θ_1, θ_2, θ_3 are the proportions of AA, Aa, and aa, respectively. This hypothesis is geometrically represented in Fig 2.

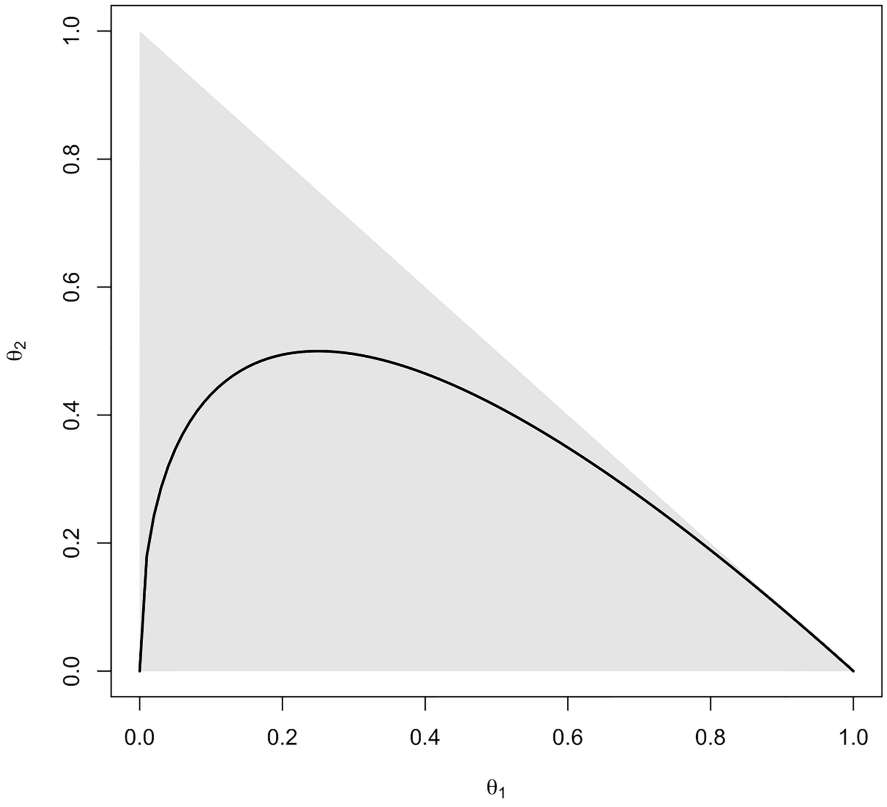


Fig 2. Geometric representation of the Hardy-Weinberg equilibrium hypothesis (black line), and the parametric space (gray shading).

https://doi.org/10.1371/journal.pone.0199102.g002

Table 3. Genotype frequency.

	AA	*Aa*	*aa*	total
X	x_1	x_2	x_3	n

$n = x_1 + x_2 + x3$.

https://doi.org/10.1371/journal.pone.0199102.t003

Let X be a random vector. Table 3 represents the genotype frequencies for the population in question. Considering n known, we assume that X follows a Trinomial(n, θ_1, θ_2, θ_3) distribution. The likelihood function for this model is

$$L(\boldsymbol{\theta} \mid \boldsymbol{x}) = \left[\frac{n!}{x_1!x_2!x_3!}\right]\theta_1^{x_1}\theta_2^{x_2}\theta_3^{x_3}, \tag{25}$$

in which $\boldsymbol{x} = \{x_1, x_2, x_3\}$, $\theta_1 + \theta_2 + \theta_3 = 1$ and $\theta_i > 0$, $i = 1, 2, 3$. Under the hypothesis $\boldsymbol{H}$,

$$L(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{H}) = \left[\frac{n!}{x_1!x_2!x_3!}\right]2^{x_2}\theta^{2x_1+x_2}(1-\theta)^{2x_3+x_2}, 0 \leq \theta \leq 1. \tag{26}$$

The maximum likelihood estimator for θ under $\boldsymbol{H}$ is $\hat{\theta} = (2x_1 + x_2)/(2n)$ and the LRT λ statistic is

$$\lambda(\boldsymbol{x}) = \frac{2^{x_2}\hat{\theta}^{2x_1+x_2}(1-\hat{\theta})^{2x_3+x_2}}{\left(\frac{x_1}{n}\right)^{x_1}\left(\frac{x_2}{n}\right)^{x_2}\left(\frac{x_3}{n}\right)^{x_3}}. \tag{27}$$

• Exact LRT p-value:

Calculations follow as for the other indices and in this scenario

$$h(\boldsymbol{x}) = \frac{n!2^{x_2}(2x_1 + x_2)!(2x_3 + x_2)!}{x_1!x_2!x_3!(2n+1)!}. \tag{28}$$

• Barnard's Exact Test:

The critical region is $\boldsymbol{R} = \{\lambda(\boldsymbol{X}) \leq \lambda(\boldsymbol{x})\}$, and the Barnard's exact p-value is obtained by

$$\text{p-value} = \max_{0\leq\theta\leq1}\sum_{\boldsymbol{R}}\left[\frac{n!}{x_1!x_2!x_3!}\right]2^{x_2}\theta^{2x_1+x_2}(1-\theta)^{2x_3+x_2}. \tag{29}$$

• FBST:

Assuming a Dirichlet(1, 1, 1) prior for $\boldsymbol{\theta}$ and that X follows a Trinomial(n, θ_1, θ_2, θ_3) distribution, the posterior distribution is $\boldsymbol{\theta} \mid \boldsymbol{x} \sim$ Dirichlet($x_1 + 1$, $x_2 + 1$, $x_3 + 1$). In this setting,

$$\sup_{\boldsymbol{\theta}\in\Theta_H}\pi(\boldsymbol{\theta} \mid \boldsymbol{x}) = \frac{(n+2)!}{x_1!x_2!x_3!}2^{x_2}\left(\frac{2x_1 + x_2}{2n}\right)^{2x_1+x_2}\left(\frac{x_2 + 2x_3}{2n}\right)^{x_2+2x_3}. \tag{30}$$

• **Other indices:**

Both asymptotic LRT p-value and asymptotic e-value are obtained, the p-value considering that $-2\ln(\lambda(\boldsymbol{X}))$ follows a X^2 distribution with 1 degrees of freedom and the FBST considering that it follows a X^2 distribution with 2 degrees of freedom.

2 Results

2.1 Relations between the indices

In many practical situations, mainly in biological studies, asymptotic distributions are used to evaluate indices even for small samples. With that in mind, one of our interests is to understand how the use of asymptotic results for small sample size settings compares to the use of an exact index. Surprisingly, the values of exact and asymptotic indexes do not diverge considerably.

As our objective is to compare the indices, we consider different scenarios for each hypothesis. For each scenario, we evaluate the significance indices of all test procedures presented here. Note that this is not a simulation study; for each sample size, we evaluate the indices for all possible contingency tables of a fixed dimension and size. For example, considering homogeneity hypothesis in a 2×2 table with marginals (10, 10), there are 121 possible tables or considering independence hypothesis in a 2×3 table with marginal 15, there are 15504 possible tables. We evaluated the indices for all tables that fit into each specification. For the e-value computation, non-informative priors for the parameters are considered (that is, $\pi(\boldsymbol{\theta}) \propto 1$). This way, no extra information is added besides the data, allowing fair comparisons between frequentist and Bayesian indices.

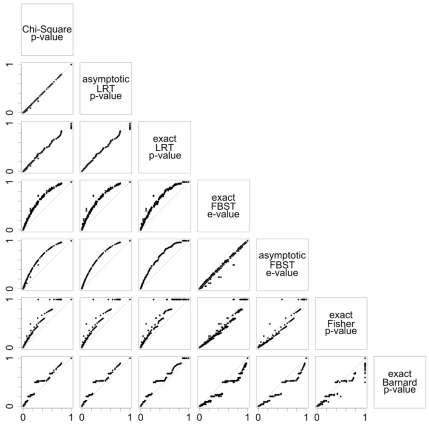
For each scenario, plots are drawn to illustrate differences between the indices' values. The indices studied are the exact LRT p-value, asymptotic p-value for the LRT, asymptotic p-value for the chi-square test, e-value and asymptotic e-value. For the homogeneity hypothesis in 2×2 tables, Fisher and Barnard exact tests were also considered, and for Hardy-Weinberg equilibrium hypothesis the Barnard's exact test was also obtained. We considered many different scenarios, however, since the aim is to understand the indices in small sample size, the scenarios presented here are in Table 4.

Figs 3, 4 and 5 illustrate the results of the discussion above. For all hypotheses, exact and asymptotic e-values are very similar for both large and small sample sizes. Looking into the frequentist indices, exact LRT p-values and asymptotic p-values, both LRT and Chi-Square, are also very similar to each other. The difference found between e-values when compared to

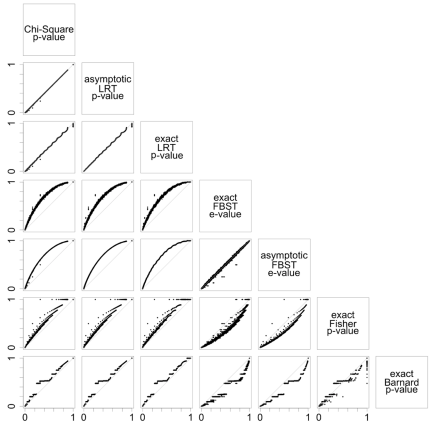
Table 4. Considered scenarios.

Setting	Hypothesis	Table	Sample sizes
1	Homogeneity	2×2	(30, 30)
2	Homogeneity	2×2	(100, 100)
3	Homogeneity	2×3	(30, 30)
4	Homogeneity	3×3	(15, 15, 15)
5	Independence	2×2	30
6	Independence	2×3	30
7	Independence	3×3	15
8	Independence	3×3	25
9	Hardy-Weinberg equilibrium	-	30
10	Hardy-Weinberg equilibrium	-	100

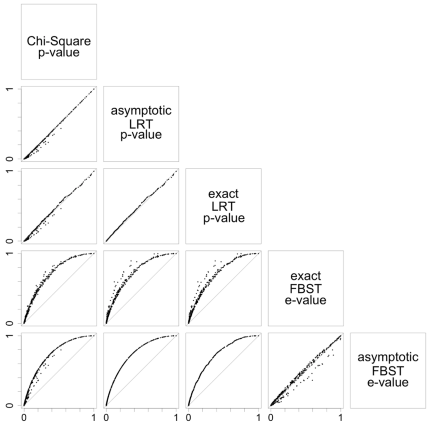
https://doi.org/10.1371/journal.pone.0199102.t004



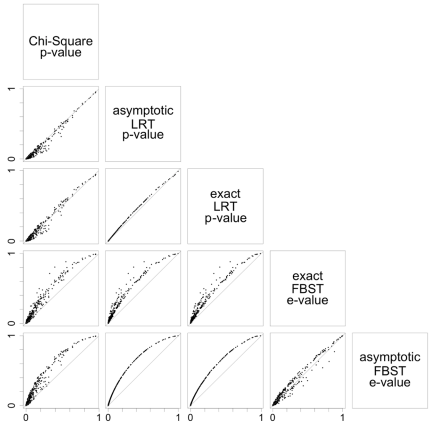
(a) 2 x 2 table, n1. = 30 and n2. = 30



(b) 2 x 2 table, n1. = 100 and n2. = 100



(c) 2 x 3 table, n1. = 30 and n2. = 30

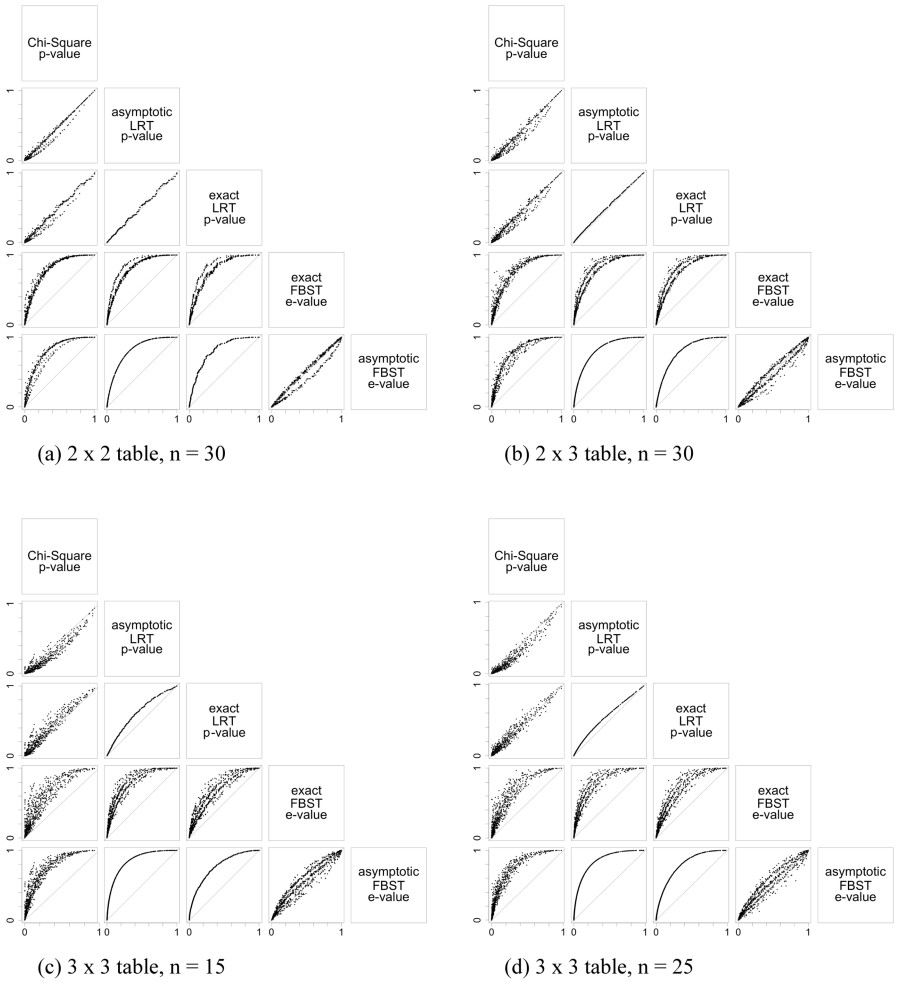


(d) 3 x 3 table, n1. = n2. = n3. = 15

Fig 3. Scaterplots for the significance indices of homogeneity hypothesis considering different sample sizes and different table dimensions. The indices were evaluated for all possible samples in the sample space. The label in the top box of that column give the index in the x-axis, and the label in the left box of that row give the index in the y-axis. Each table dimesions and sample sizes are given in the sublabels.

https://doi.org/10.1371/journal.pone.0199102.g003

asymptotic LRT p-value happens as a result of the way these indices are formulated: while e-values consider the full dimension of the parameter space, p-value consider the complementary dimension of the set corresponding to hypothesis $\boldsymbol{H}$. This is expected from the asymptotic relationship between e-value and p-value from the LRT [13, 14]. Since the exact LRT p-value is directly related to the asymptotic LRT p-value, we observe the same behavior of the differences between e-values and asymptotic LRT p-value. Fisher's exact test was only calculated for the homogeneity hypothesis in 2×2 tables, and Barnard's exact test was calculated for the homogeneity hypothesis in 2×2 tables and for the Hardy-Weinberg equilibrium hypothesis. Both indices have a different behavior among the other indices considered. They have a discrete behavior, which is not surprising since Fisher's exact test is a conditional test and Barnard's exact test takes a maximization nuisance parameter elimination. Looking at the plots, their values do not form a continuous curve like the other indices' values do, and its points are quite far from all the other indices.



(a) 2 x 2 table, n = 30

(b) 2 x 3 table, n = 30

(c) 3 x 3 table, n = 15

(d) 3 x 3 table, n = 25

Fig 4. Scaterplots for the significance indices of independence hypothesis considering different sample sizes and different table dimensions. The indices were evaluated for all possible samples in the sample space. The label in the top box of that column give the index in the x-axis, and the label in the left box of that row give the index in the y-axis. Each table dimesions and sample sizes are given in the sublabels.

https://doi.org/10.1371/journal.pone.0199102.g004

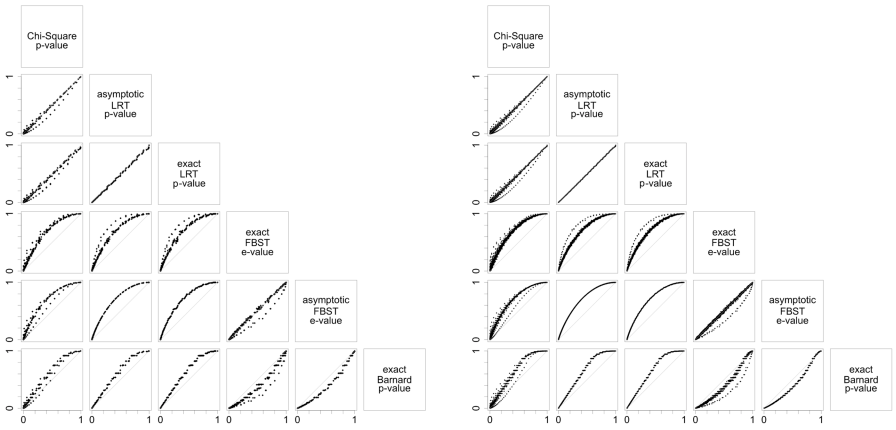
2.2 Power function

Power functions are a useful tool to compare hypothesis tests. For all $\theta \in \Theta$, the power function provides the probability of rejecting the hypothesis for a given θ. In fact, we look for a test that does not reject the hypothesis for $\theta \in \Theta_{\boldsymbol{H}}$ and the further the θ value is from the hypothesis, the probability of rejection increases.

The power functions presented are the ones that we are able to represent in $\mathbb{R}^3$, which are the power functions for the homogeneity hypothesis in 2×2 contingency tables and for the Hardy-Weinberg equilibrium hypothesis.

We used p-values less than 0.05 as a decision rule to reject the hypothesis. This choice is based on what is vastly used in most fields of science as a decision rule. In this case, Power (θ_1, θ_2) = P(reject $\boldsymbol{H}|(\theta_1, \theta_2)$ and Reject $\boldsymbol{H}$ if index ≤ 0.05.

We obtain the power function for all tests but the FBST. The FBST is a Bayesian significance test and in order to obtain a power function, one would need a decision rule. Since its



(a) n = 30 (b) n = 100

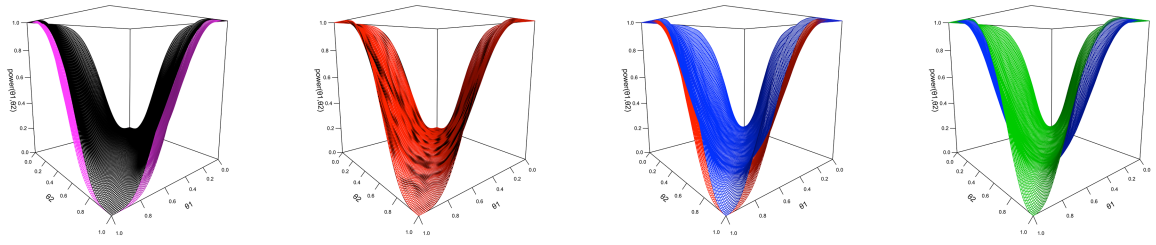
Fig 5. Scaterplots for the significance indices of Hardy-Weinberg hypothesis considering different sample sizes and different table dimensions. The indices were evaluated for all possible samples in the sample space. The label in the top box of that column give the index in the x-axis, and the label in the left box of that row give the index in the y-axis. Each table dimesions and sample sizes are given in the sublabels.

https://doi.org/10.1371/journal.pone.0199102.g005

construction differs from that of the p-values, we cannot use the same decision rule, and constructing a decision rule is not in the scope of this paper.

We used a Monte Carlo procedure to evaluate the power function of these tests. We consider a grid for the unit square with 100 × 100 points on the axes (θ_1, θ_2). For each point in the grid we generated 1000 tables. From these 1000 tables we evaluate the proportion of rejections, which is an approximation of the power function.

We plot pairs of power functions to illustrate and compare their shapes. For the homogeneity hypothesis in a table with marginals (10, 10), Fig 6 shows that Fisher's exact test is less powerful than the Barnard's exact test, the Barnard's exact test is has similar power when compared with the Chi-square test, while the Chi-square is less powerful than the proposed exact LRT p-value, which is less powerful than the asymptotic p-value for the LRT. To have a clear picture, we plot the power functions from different tests against each other. Fig 7a consists of the power functions for tables with marginal equals to (10, 10). It shows that the use of the asymptotic p-value for the LRT results in a more powerful test than the other indices. When comparing the proposed exact p-value to other indices, it is more powerful than the Chi-square test and the Fisher's exact test. Between the Chi-square and the Fisher's exact test, the Chi-square test is more powerful.



(a) Pink: power function considering the Fisher's exact test; and black: power function considering the Barnard's exact test.

(b) Black: power function considering the Barnard's exact test; and red: power function considering the Chi-square teste.

(c) Red: power function considering the Chi-square test; and blue: power function considering the exact LRT test (p-value).

(d) Blue: power function considering the exact LRT test (p-value); and green: the asymptotic LRT test (p-value).

Fig 6. Power function for homogeneity hypothesis in 2 × 2 contingency tables with $n_{1\cdot} = n_{2\cdot} = 10$.

https://doi.org/10.1371/journal.pone.0199102.g006

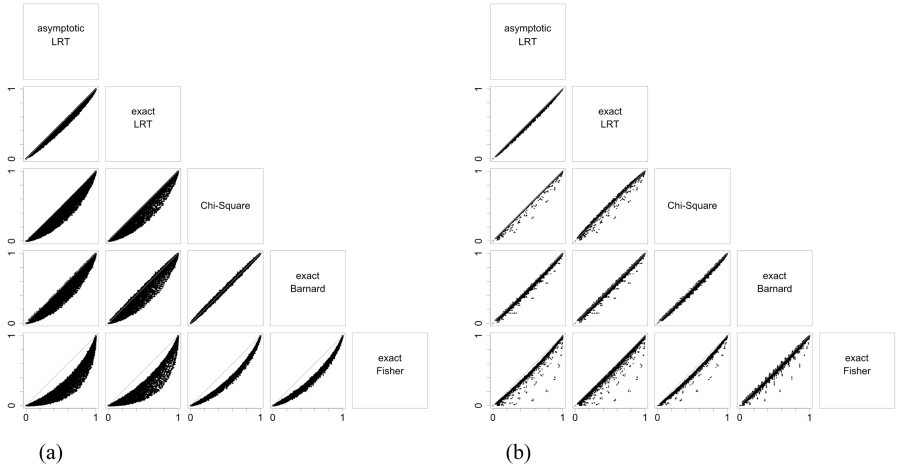


Fig 7. Plots of power function values for the homogeneity test. Each graph presents one index versus another, each dot representing a point in the considered parametric space (in this case, 100 × 100 = 10000 points), and if a dot is on top of the gray identity line, the power functions assume the same value for that point in the parametric space. The scenario is 2 × 2 with marginals $n_{1\cdot} = n_{2\cdot} = 10$ in (a) and $n_{1\cdot} = n_{2\cdot} = 100$ in (b).

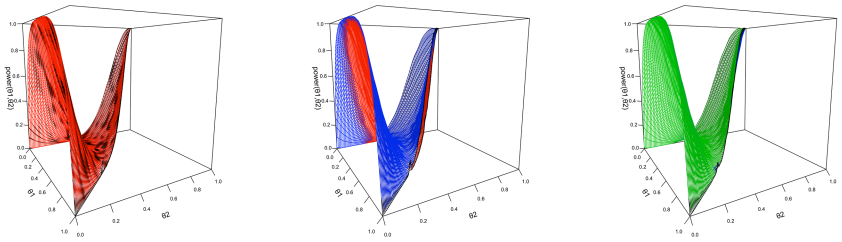
https://doi.org/10.1371/journal.pone.0199102.g007

For tables with marginal equals to (100, 100), the graphs are more concentrated near the identity line (Fig 7b), showing that all indices are more alike. The ordering still exists, but it is less severe. It is interesting to point out that, as expected, the Chi-square test works better with larger samples.

For the Hardy-Weinberg hypothesis, the results are similar to the ones obtained for the homogeneity hypothesis and are shown in Figs 8 and 9. In this case, the most powerful test was the asymptotic p-value for the LRT, followed by the exact p-value for the LRT, which is more powerful to the Chi-square test, that is similar the Barnard's exact test. We call attention to the fact that, under hypothesis $\boldsymbol{H}$, the power function achieves the value of 0.05, as expected, since this is the significance level chosen to build the power functions.

3 Conclusion

After evaluating the indices for tables in different scenarios, we noticed that all of them had very similar behaviors, independently of the perspective (Bayesian or frequentist), sample size and table dimension. The exceptions are the p-values for Fisher and Barnard's exact tests for the homogeneity hypothesis in 2 × 2 tables, and Barnard's exact test for Hardy-Weinberg



(a) Black: power function considering the Barnard's exact test; and red: power function considering the Chi-square test.

(b) Red: power function considering the Chi-square test; and blue: power function considering the exact LRT test (p-value).

(c) Blue: power function considering the exact LRT test (p-value); and power function considering the asymptotic LRT test (p-value).

Fig 8. Power function for Hardy-Weinberg equilibrium hypothesis with $n = 10$.

https://doi.org/10.1371/journal.pone.0199102.g008

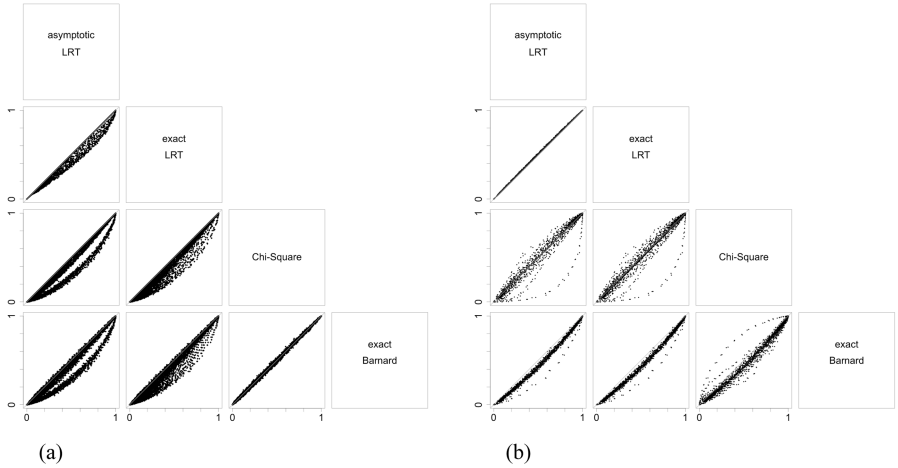


Fig 9. Plots of power functions values for the Hardy-Weinberg equilibrium test. Each graph presents one index versus another, each dot representing a point in the considered parametric space (in this case, $100 \times 100 = 10000$ points), and if a dot is on top of the gray identity line, the power functions assume the same value for that point in the parametric space. The scenarios are marginals $n = 10$ (a) and $n = 100$ (b).

https://doi.org/10.1371/journal.pone.0199102.g009

equilibrium, which show a discretized behavior. Studying the power functions considering homogeneity hypothesis in 2×2 tables and Hardy-Weinberg equilibrium hypothesis, the LRT presented itself as a powerful test when considering small sample sizes, while Fisher's exact test was the least powerful one for the homogeneity hypothesis and the Barnard's exact test was the least powerful for the Hardy-Weinberg equilibrium hypothesis. By enlarging sample sizes, the power of these tests increases accordingly.

Finally, we finish this paper listing our main conclusions:

- The LTR asymptotic p-value seems to be a good frequentist alternative for small sample sizes.
- Since there is an asymptotic relationship between the p-value for the LRT and the e-value (FBST), we consider that both indices are equivalent in the explored settings.
- In cases where there is available information besides the data that to be taken into account, represented by informative priors, we consider the e-value a more appropriate index than a frequenstist one, since the e-value offers a mechanism to incorporate that information.

Acknowledgments

This work was partially supported by the Brazilian agencies FAPESP grant 2012/16669-4, and CNPq grants 302767/2017-7 and 308776/2014-3. The agencies had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

Conceptualization: Natalia L. Oliveira, Carlos A. de B. Pereira, Marcio A. Diniz, Adriano Polpo.

Formal analysis: Natalia L. Oliveira, Carlos A. de B. Pereira, Marcio A. Diniz, Adriano Polpo.

Methodology: Natalia L. Oliveira, Carlos A. de B. Pereira, Marcio A. Diniz, Adriano Polpo.

Supervision: Adriano Polpo.

Writing – original draft: Natalia L. Oliveira, Carlos A. de B. Pereira, Marcio A. Diniz, Adriano Polpo.

Writing – review & editing: Natalia L. Oliveira, Carlos A. de B. Pereira, Marcio A. Diniz, Adriano Polpo.

References

1. Emigh TH. A Comparison of Tests for Hardy-Weinberg Equilibrium. Biometrics. 1980; 36(4):627–642. https://doi.org/10.2307/2556115 PMID: 25856832
2. Montoya-Delgado LE, Z IT, Pereira CAB, Whittle MR. An unconditional exact test for the Hardy-Weimberg Equilibrium Law: Sample space ordering using the Bayes Factor. Genetics. 2001; 158(2):875–83. PMID: 11404348
3. Pereira CAB, Wechsler S S. On the Concept of P-value. Brazilian Journal of Probability and Statistics. 1993; 7:159–177.
4. Wilks SS. The large-sample distribution of the likelihood ratio for testing composite hypotheses. The Annals of Mathematical Statistics. 1938; 9:60–62. https://doi.org/10.1214/aoms/1177732360
5. Fisher RA. Statistical Methods for Research Workers. 5th ed. Biological Monographs and Manuals. Edinburg: Oliver and Boyd; 1934.
6. Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, Series 5. 1900; 50(302):157–175. https://doi.org/10.1080/14786440009463897
7. Fisher RA. On the interpretation of χ^2 from contingency tables, and the calculation of P. Journal of the Royal Statistical Society. 1922; 85(1):87–94. https://doi.org/10.2307/2340521
8. Barnard GA. A New Test for 2x2 Tables. Nature. 1945; 156:177. https://doi.org/10.1038/156177a0
9. Fisher RA. A New Test for 2x2 Tables. Nature. 1945; 156:388. https://doi.org/10.1038/156388a0
10. Barnard GA. A New Test for 2x2 Tables. Nature. 1945; 156:783–784. https://doi.org/10.1038/156177a0
11. Barnard GA. Statistical Inference. Journal of the Royal Statistical Society, Series B (Methodological). 1949; 11(2):115–149.
12. Pereira CAB, Stern JM. Evidence and Credibility: a Full Bayesian Test of Precise Hypothesis. Entropy. 1999; 1:104–115.
13. Pereira CAB, Stern JM, Wechsler S. Can a Significance Test Be Genuinely Bayesian? Bayesian Analysis. 2008; 3(1):19–100. https://doi.org/10.1214/08-BA303
14. Diniz MA, Pereira CAB, Polpo A, Stern J, Wechesler S. Relationship Between Bayesian and Frequentist Significance Indices. International Journal for Uncertainty Quantification. 2012; 2(2):161–172. https://doi.org/10.1615/Int.J.UncertaintyQuantification.2012003647
15. Wasserstein RL, Lazar NA. The ASA's Statement on p-Values: Context, Process, and Purpose. The American Statistician. 2016; 70(2):129–133. https://doi.org/10.1080/00031305.2016.1154108
16. Lawson AE, Clark B, Cramer-Meldrum E, Falconer KA, Sequist JM, Kwon Y. Development of Scientific Reasoning in College Biology: Do Two Levels of General Hypothesis-Testing Skills Exist? Journal of Research in Science Teaching. 2000; 37(1):81–101. https://doi.org/10.1002/(SICI)1098-2736(200001)37:1%3C81::AID-TEA6%3E3.0.CO;2-I
17. Herrmann E, Call J, Hernandez-Lloreda MV, Hare B, Tomasello M. Humans Have Evolved Specialized Skills of Social Cognition: The Cultural Intelligence Hypothesis. Science. 2007; 317(5843):1360–1366. https://doi.org/10.1126/science.1146282 PMID: 17823346
18. Montgomery DD, Runger GC. Applied Statistics and Probability for Engineers. John Wiley & Sons; 2010.
19. Agresti A. Exact inference for categorical data: recent advances and continuing controversies. Statistics in Medicine. 2001; 20:2709–2722. https://doi.org/10.1002/sim.738 PMID: 11523078
20. Agresti A. Categorical Data Analysis. 2nd ed. John Wiley & Sons; 2002.
21. Mehta CR, F HJ. Exact Power of Conditional and Unconditional Tests: Going beyond the 2x2 Contingency Table. The American Statistician. 1993; 47(2):91–98. https://doi.org/10.1080/00031305.1993.10475946

22. Eberhardt KR, Fligner MA. A Comparison of Two Tests for Equality of Two Proportions. The American Statistician. 1977; 31(4):151–155. https://doi.org/10.1080/00031305.1977.10479225
23. Pagano M, Halvorsen KT. An Algorithm for Finding the Exact Significance Levels of r × c Contingency Tables. Journal of the American Statistical Association. 1981; 76(376):931–934. https://doi.org/10.2307/2287590
24. Irony TZ, Pereira CAB, Tiwari RC. Analysis of Opinion Swing: Comparison of two correlated proportions. The American Statistician. 2000; 54(1):57–62. https://doi.org/10.2307/2685613
25. Zhang L, Xinzhong Xu, Chen G. The Exact Likelihood Ratio Test for Equality of Two Normal Populations. The American Statistician. 2012; 66(3):180–184. https://doi.org/10.1080/00031305.2012.707083
26. Shan G, Wilding GE. Powerful Exact Unconditional Tests for Agreement Between Two Raters with Binary Endpoints. PLoS ONE. 2014; 9(5):e97386. https://doi.org/10.1371/journal.pone.0097386 PMID: 24837970
27. Basu D. On the Elimination of Nuisance Parameters. Journal of the American Statistical Association. 1977; 72(358):355–366. https://doi.org/10.1080/01621459.1977.10481002
28. Casella G, Berger R. Statistical Inference. 2nd ed. Duxbury Press; 2001.
29. Agresti A. An Introduction to Categorical Data Analysis. 2nd ed. John Wiley & Sons; 2007.
30. Irony TZ, Pereira CAB. Exact tests for equality of two proportions: Fisher vs. Bayes. Journal of Statistical Computation and Simulation. 1986; 25:93–114. https://doi.org/10.1080/00949658608810926
31. Hartl DL, Clark AG. Principles of Population Genetics. 4th ed. Sinauer Associates, Inc. Publishers; 2007.